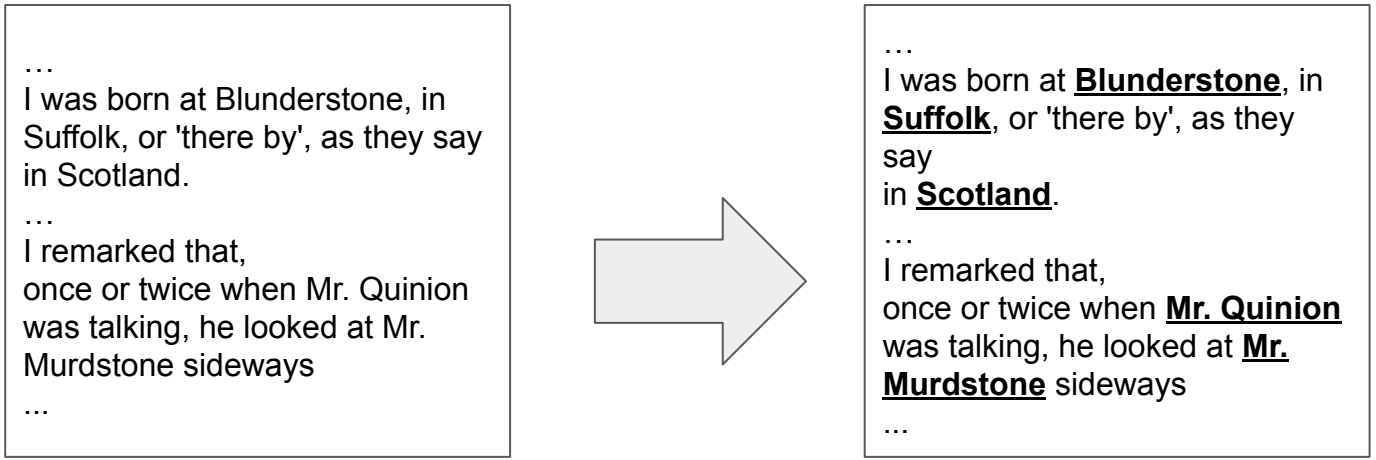
Les entités nommées (et au-delà) comme outil d'analyse de corpus textuels : une perspective de projet en humanités numériques

Carmen Brando (PGHN / CRH, EHESS)
carmen.brando@ehess.fr

Séminaire du PIREH - Paris, le 20 octobre 2023

Les entités nommées (EN) en traitement automatique des langues (TAL)

Extrait de David Copperfield de Charles Dickens



texte brut

Texte annoté en EN

Les EN en TAL

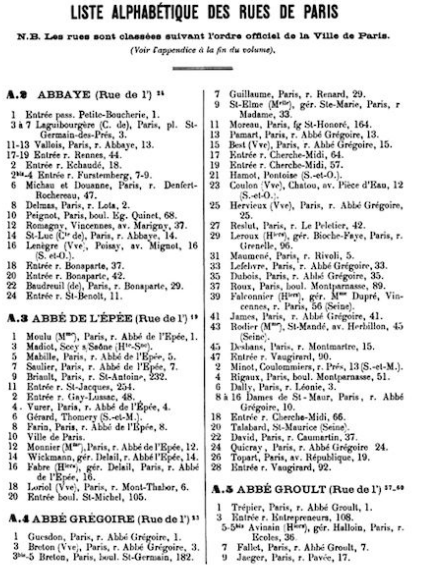
- "*Il s'agit de types d'unités lexicales particulières qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques, et qui est désignée par son nom*" (Ester)
- associée aux **noms propres** (*Barack Obama*, *Hugo, rue Mouffetard*) et parfois à des **descriptions définies** (un groupe nominal, ex : *le chat noir*)
- Les trois **catégories** (dites *coarse-grained*) : noms de personnes, de lieux, d'organisations ; extension de la typologie d'EN pour inclure : les fonctions de personnes, les productions humaines, les expressions numériques, temporelles, etc.

Les EN en TAL: types d'ambigüités

- le même nom est utilisé pour plusieurs entités
  - *Paris* (France) et *Paris* (Texas)
- une même entité peut avoir plusieurs noms
  - *Paris*, *Paname*
- le nom d'une entité peut être utilisé pour désigner une entité en catégories différentes (le cas de la métonymie)
  - la *Sorbonne*, la *France*, … une organisation mais peut être aussi un lieu selon le contexte
  - "Le *prix Nobel de la Paix* s'est montré digne devant une telle épreuve"

Structurer et enrichir des corpus numériques/numérisés : avec les EN et au delà

Annuaires des propriétaires et de propriétés de Paris (PTM)



LISTE ALPHABÉTIQUE DES RUES DE PARIS

N.B. Les rues sont classées suivant l'ordre officiel de la Ville de Paris.

(Voir l'appendice à la fin du volume).

A.2 ABBAYE (Rue de l') [11]

1 Entrée pass. Petite-Boucherie, 1.
3 à 7 Laguibourgère (C. de), Paris, pl. St-Germain-des-Près, 3.
11-13 Vallais, Paris, r. Abbaye, 13.
17-19 Entrée r. Rennes, 44.
2 Entrée r. Echaudé, 18.
2bis-4 Entrée r. Furstenberg, 7-9.
6 Michau et Douanne, Paris, r. Denfert-Rochereau, 47.
8 Delmas, Paris, r. Lota, 2.
10 Peignot, Paris, boul. Ep. Quinet, 68.
12 Romagny, Vincennes, av. Marigny, 37.
14 St-Luc (Cte de), Paris, r. Abbaye, 14.
16 Lenigre (Vve), Poissy, av. Mignot, 16 (S. et-O.).
18 Entrée r. Bonaparte, 37.
20 Entrée r. Bonaparte, 42.
22 Baudreuil (de), Paris, r. Bonaparte, 29.
24 Entrée r. St-Benoît, 11.

A.3 ABBÉ DE L'ÉPÉE (Rue de l') [12]

1 Moulu (Mme), Paris, r. Abbé de l'Epée, 1.
3 Madiot, Scey s/Seine (Hte-Se).
5 Mabille, Paris, r. Abbé de l'Epée, 5.
7 Saulier, Paris, r. Abbé de l'Epée, 7.
9 Brisult, Paris, r. St-Antoine, 232.
11 Entrée r. St-Jacques, 254.
2 Entrée r. Guy-Lussac, 48.
4 Vurer, Paris, r. Abbé de l'Epée, 4.
6 Gérard, Thomery (S.-et-M.).
8 Paris, Paris, r. Abbé de l'Epée, 8.
10 Ville de Paris.
12 Monnier (Mme), Paris, r. Abbé de l'Epée, 12.
14 Wickmann, gér. Delail, r. Abbé l'Epée, 14.
16 Fabre (Hte), gér. Delail, Paris, r. Abbé de l'Epée, 16.
18 Loriol (Vve), Paris, r. Mont-Thabor, 6.
20 Entrée boul. St-Michel, 105.

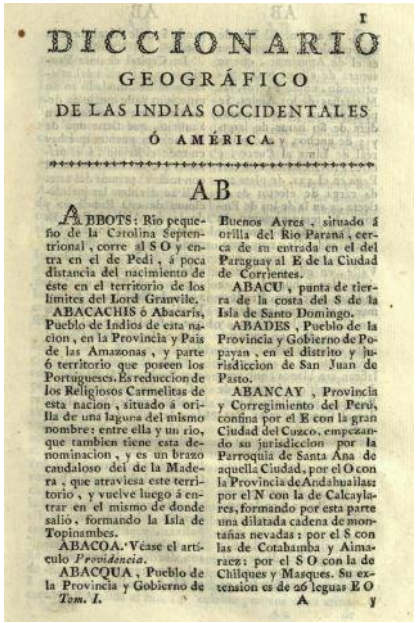
A.4 ABBÉ GRÉGOIRE (Rue de l') [13]

1 Gueudon, Paris, r. Abbé Grégoire, 1.
3 Breton (Vve), Paris, r. Abbé Grégoire, 3.
3bis-5 Breton, Paris, boul. St-Germain, 182.
7 Guillaume, Paris, r. Renard, 29.
9 St-Elme (Mlle), gér. Ste-Marie, Paris, r Madame, 33.
11 Moreau, Paris, fg St-Honoré, 164.
13 Pamart, Paris, r. Abbé Grégoire, 13.
15 Best (Vve), Paris, r. Abbé Grégoire, 15.
17 Entrée r. Cherche-Midi, 64.
19 Entrée r. Cherche-Midi, 57.
21 Hamot, Pontoise (S.-et-O.).
23 Coulon (Vve), Chatou, av. Pièce d'Eau, 12 (S.-et-O.).
25 Hervieux (Vve), Paris, r. Abbé Grégoire, 25.
27 Reslut, Paris, r. Le Pelletier, 42.
29 Leroux (Hte), gér. Bioche-Faye, Paris, r. Grenelle, 96.
31 Maumenet, Paris, r. Rivoli, 5.
33 Lefelvre, Paris, r. Abbé Grégoire, 33.
35 Dubois, Paris, r. Abbé Grégoire, 35.
37 Roux, Paris, boul. Montparnasse, 89.
39 Falconnier (Hte), gér. Mme Dupré, Vincennes, r. Paris, 56 (Seine).
41 James, Paris, r. Abbé Grégoire, 41.
43 Bodier (Mme), St-Mandé, av. Herblillon, 45 (Seine).
45 Deshans, Paris, r. Montmartre, 15.
47 Entrée r. Vaugirard, 90.
2 Minot, Coulommiers, r. Prés, 13 (S.-et-M.).
4 Rigaux, Paris, boul. Montparnasse, 51.
6 Dally, Paris, r. Léonie, 3.
8 à 16 Dames de St-Maur, Paris, r. Abbé Grégoire, 10.
18 Entrée r. Cherche-Midi, 66.
20 Talabard, St-Maurice (Seine).
22 David, Paris, r. Caumartin, 37.
24 Quicray, Paris, r. Abbé Grégoire 24.
26 Topart, Paris, av. République, 19.
28 Entrée r. Vaugirard, 92.

A.5 ABBÉ GROULT (Rue de l') [27-49]

1 Trépier, Paris, r. Abbé Groult, 1.
3 Entrée r. Entrepreneurs, 108.
5-5bis Avinain (Hte), gér. Halloin, Paris, r. Ecoles, 36.
7 Fallet, Paris, r. Abbé Groult, 7.
9 Jaeger, Paris, r. Pavée, 17.

Dictionnaire géo-historique d'Alcedo sur l'Amérique hispanique (ANR TopUrbi)



I

DICCIONARIO

GEOGRÁFICO

DE LAS INDIAS OCCIDENTALES

Ó AMÉRICA.

AB

ABBOTS: Rio pequeño de la Carolina Septentrional, corre al S O y entra en el de Pedi, á poca distancia del nacimiento de este en el territorio de los límites del Lord Granville.

ABACACHIS ó Abacaris, Pueblo de Indios de esta nacion, en la Provincia y Pais de las Amazonas, y parte ó territorio que poseen los Portugueses. Es reduccion de los Religiosos Carmelitas de esta nacion, situado á orilla de una laguna del mismo nombre: entre ella y un rio, que tambien tiene esta denominacion, y es un brazo caudaloso del de la Madera, que atraviesa este territorio, y vuelve luego á entrar en el mismo de donde salió, formando la Isla de Topinambes.

ABACOA. Véase el artículo *Providencia*.

ABACQUA, Pueblo de la Provincia y Gobierno de Buenos Ayres, situado á orilla del Rio Paraná, cerca de su entrada en el del Paraguay al E de la Ciudad de Corrientes.

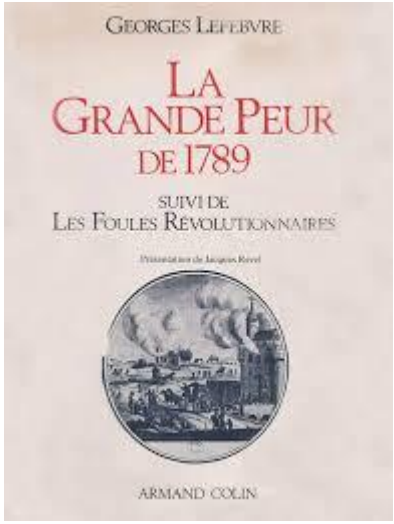
ABACU, punta de tierra de la costa del S de la Isla de Santo Domingo.

ABADES, Pueblo de la Provincia y Gobierno de Popayan, en el distrito y jurisdiccion de San Juan de Pasto.

ABANCAY, Provincia y Corregimiento del Perú, confina por el E con la gran Ciudad del Cuzco, empezando su jurisdiccion por la Parroquia de Santa Ana de aquella Ciudad, por el O con la Provincia de Andahuailas: por el N con la de Calcaylares, formando por esta parte una dilatada cadena de montañas nevadas: por el S con las de Cotahamba y Aimaraez: por el S O con la de Chilques y Masques. Su extension es de 26 leguas E O

Tom. I. A y

La grande peur de 1789 de Georges Lefebvre



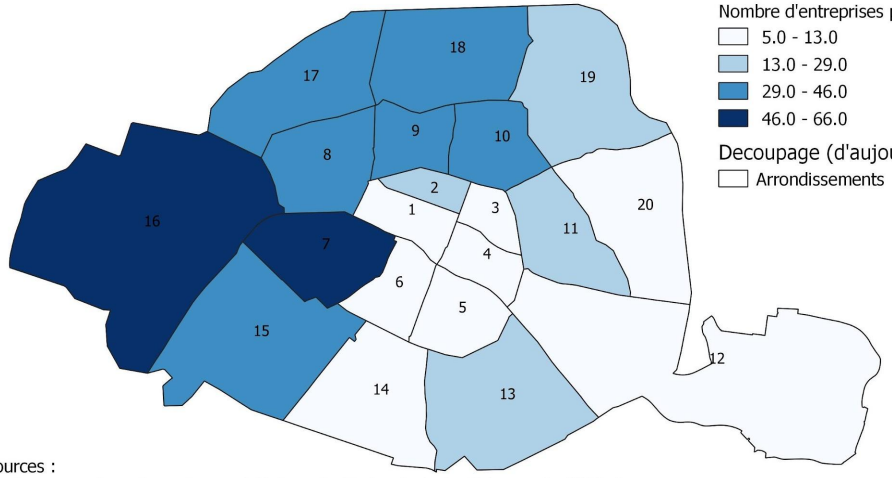
Articles scientifiques en anthropologie et histoire des savoirs (*Savoirs*)

Applications des EN dans les HN

- Cartographier et visualiser les EN d’un corpus, par exemple, une cartographie ou analyse de la dimension spatiale et spatio temporelle d’un corpus
- Interroger de manière croisée des corpus et bases de données différents reliés par des EN
- Indexer pour préparer une édition numérique enrichie
- Calculer de suggestions de lecture de textes dans un corpus

PTM Annuaires, work in progress (référentiel géohistorique en construction)

La propriété des entreprises en 1898 à Paris



Adresses de propriétés (localisation d'aujourd'hui)

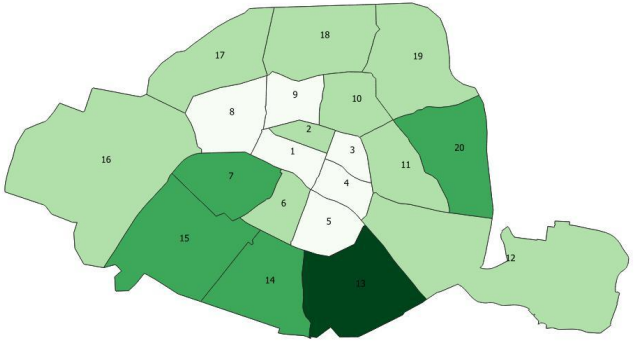
Nombre d'entreprises par arrondissement

- 5.0 - 13.0
- 13.0 - 29.0
- 29.0 - 46.0
- 46.0 - 66.0

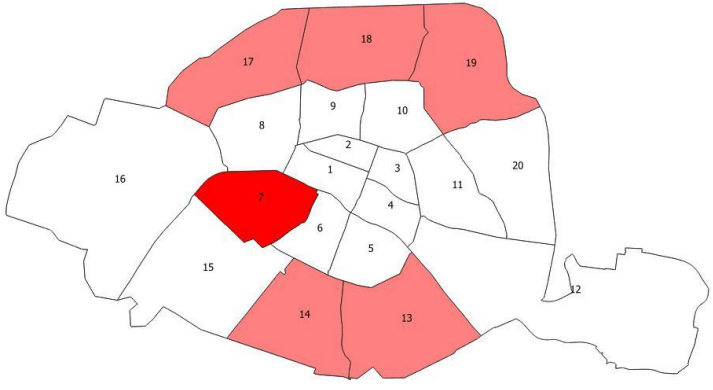
Decoupage (d'aujourd'hui)

- Arrondissements

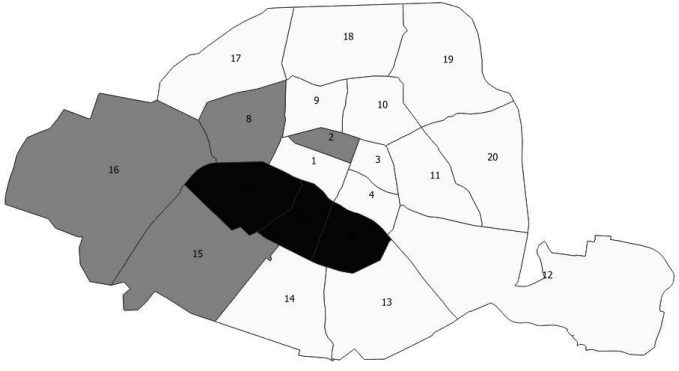
Sources :
- adresses : Annuaires de propirétaires de Paris et de la Seine, vol. 1898. Les localisations sont issues de https://geo.api.gouv.fr/adresse.
- decoupage administratif : Ville de Paris (https://opendata.paris.fr/)



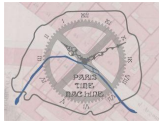
par l'Etat



par la compagnie de Chemins de Fer



par l'Eglise catholique



Mise en oeuvre de chaînes de traitement des sources aux analyses

- tout en intégrant des approches issues des **différents domaines** : TAL, analyse des données, géomatique
- plusieurs approches/outils et solutions partielles existent mais l'enjeu est de faciliter (**généraliser**) leur **réutilisation** dans des projets avec des besoins proches
- Adapter la pipeline selon les **niveaux de structuration des textes** ; les EN, une composante de la pipeline
- Produire des données **réutilisables et exploitables** pour des recherches, les enrichir et spatialiser
- développer et **systématiser** des analyses à partir de ces données
- Pipeline EN/TAL vs pipeline HN : objectifs proches mais différents (unitâche vs. multitâche)

Mise en oeuvre d’une chaîne de traitement des sources aux analyses : sur les outils

- performance des outils
- robustesse et généricité des outils
- outils en production avec une communauté et un appui technique
- interfaces utilisateurs accessibles à des utilisateurs non experts
- formats standards d’échange divers et variés

68-70 Hurand, Paris, av. Ternes, 38.
72 Devauchelle et Longuet, Bouchon (Somme).
74 Chartier, Paris, r. Montmartre, 157.
76 Gillet, Paris, r. N.-D. Nazareth, 36.
78 Renault, Noisy-le-Grand (S.-et-O.).
80 Baroux, Paris, boul. Beauséjour, 43.
82 Grimbert, Paris, cité Tulma, 11.
84 Daguet (Mme), Paris, r. Malaillon, 18.
86 Bertron-Auger, La Flèche (Sarthe).
88 Calmels, Paris, boul. St-Marcel, 9.
90 Deville, Paris, r. Bruxelles, 23.
92 Decolange, Versailles, av. St-Cloud, 20.
94 Cros, Paris, r. Lombards, 23.
96 Masson (Mme), gér. Rouchon, Paris, r. Quincampoix, 96.
98 Vincent, Paris, r. Quincampoix, 98.
100 Sanoner (Vve), Paris, r. Rivoli, 130.

Q.6
QUINZE-VINGTS (Passage des) 48

1 Entrée r. Lyon, 46.
3 Davenne, Paris, pass. Quinze-Vingts, 3.
s. nº Ville de Paris.

R.1 RABELAIS (Rue) 30_31

1 Eiffel, Paris, r. Rabelais, 1.
3 Deleuze, Paris, r. Rabelais 3.
5 Entrée r. Montaigne, 26.
2-4 Gérard (Bon), Paris, fg St-Honoré, 85.
6-8 Aligre (Me d'), Paris, fg St-Honoré, 89.

R.4 RACINE (Rue) 22

1 Ferry, Paris, r. Turin, 27.
3 Levaux (Mme), Paris, r. Rennes, 108.
5 Clolus, gér. Ferrembach, Paris, r. Racine, 5.
7 Bourrières, gér. Arnould, Paris, r. Racine, 7.
9 Barrois, St-Germain-en-Laye (S.-et-O.).
11 Ville de Paris.
13 Finelle, St-Mandé, Chre Etang, 12 (Seine).
15 Gille (Mme), Paris, r. Racine, 15.
17 Chanroux, gér. Roussel, Paris, r. Racine, 17.
19 Entrée r. Monsieur le Prince, 22.
21 Normand (Vve), Paris, boul. Beaumarchais, 50.
23-25 Blondeau (Mme), Paris, av. Percier, 8bis.
2 Crochard (Mme), gér. Vve David, Paris, r. Médicis, 7.
4-6 Ville de Paris.
8 Ecole Nationale des Arts décoratifs (Etat).
s. nº Ecole de Médecine.
24 Entrée r. Monsieur le Prince, 20
26-28 Flammarion, St-Mandé, av. Daumesnil, 39 (Seine).
30 Croville, Paris, r. Monsieur le Prince, 2.

R.6 RADZIWILL (Rue) 3

9 Raincourt (Mme), St-Germain-en-Laye, r. Voltaire, 44 (S.-et-O.).
11 Blot (Vve), gér. Amelin, Paris, r. St-Pétersbourg, 24.
13 Croix, Paris, r. Puteaux, 7.
15-17 Fattet (Mme), Paris, av. Champs-Elysées, 21.
19 Bertet, Paris, r. Radziwill, 19.
21 Charles (Vve), Paris, gal. Montpensier, 25.
23 Delâttre, Paris, r. Washington, 30.
25-27 Banque de France.
29 Chevet (Vve), Paris, r. François Gérard, 43.
31 Chauvelot, gér. Viron et Provost, Paris, r. Radziwill, 31.
33-35 Banque de France.
37 Cousin, Paris, r. Boccador, 3.
2 Banque de France.

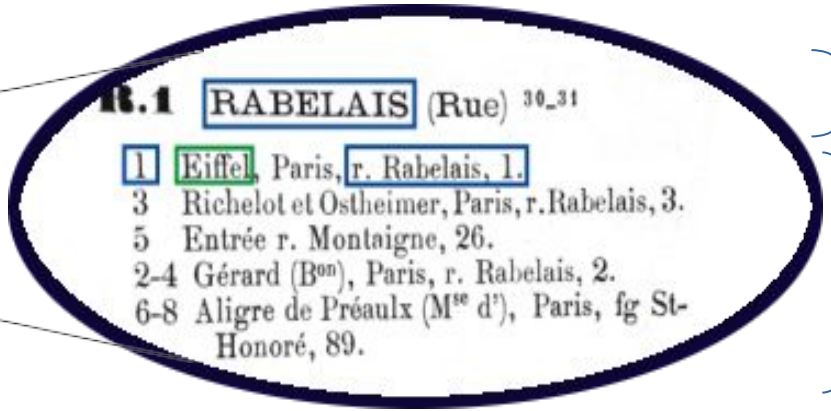
R.7 RAFFET (Rue) 61

1 Entrée r. Source, 36.
3-5 Thomas (Vve), Paris, r. Beauséjour, 49.
s. nº Ville de Paris.
19 Richy, Paris, r. Raffet, 19.
19bis Soleville, Paris, boul. Beauséjour, 4.
21 Hallet (Mme), Paris, r. Raffet, 21.
23 St-Lanne, Paris, r. Raffet, 23.
s. nº Robert, Paris, pl. Possoz, 2.
s. nº Noblet, Paris, r. Cujas, 13.
2 Prévoteau, St-Germain-en-Laye, r. Paris, 2 (S.-et-O.).
4-6 Laperche, Paris, fg St-Honoré, 118.
8 Entrée r. Jasmin, 17.
10 à 14 Heine (Bue), Paris, r. Victoire, 63.
16 Audewal et Wehry, Paris, r. Raffet, 16.

R.8 RAGUINOT (Passage) 18

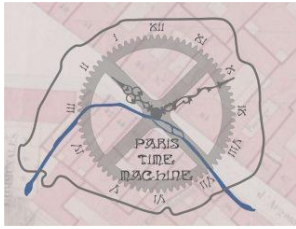
1 Entrée r. Chalon, 24.
3 Lebœuf, Alfortville (Seine).
5 Robillard, Paris, av. Parmentier, 7.
7 Dubois, gér. Rex-Combes, Paris, r. Chaligny, 15.
9 Besneux, Paris, r. Lyon, 46.
9bis Maurel, gér. Salles, Paris, pass. Raguinot, 18
11 Combaco, Paris, av. Gde-Armée, 87.
11bis Bourdat, Paris, r. Charenton, 100bis.
13 Buzenet Vve), Houilles (S.-et-O.).
15 Lavolpillière, gér. Sagnot, Paris, pass. Raguinot, 15.
17 Carrier, Pass. Raguinot, 17.
19 Lemoine, Chateau Thierry (Aisne).
21-23 Jacquetan, Nogent-s-Marne (Seine).
25 Leverrier, Paris, pass. Raguinot, 25.
27 Mailbochon, Paris, r. Lecourbe, 183.
29 Cheron, gér. Flajol, Paris, pl. Nation, 14.

Pipeline adapté selon les niveaux de structuration des textes



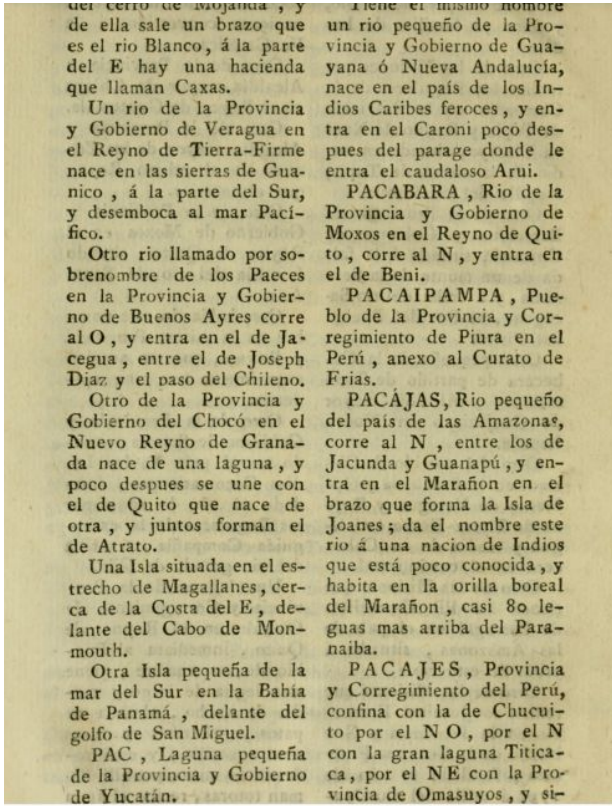
Macrostructure

Microstructure



Annuaire de propriétaires et de propriétés de Paris et de la Seine (Projet Time Machine)

Pipeline adapté selon les niveaux de structuration des textes : Dictionnaire d'Alcedo, entrée pour "Pacajes" (TopUrbi)



de ella sale un brazo que es el rio Blanco, á la parte del E hay una hacienda que llaman Caxas.

Un rio de la Provincia y Gobierno de Veragua en el Reyno de Tierra-Firme nace en las sierras de Guanico, á la parte del Sur, y desemboca al mar Pacífico.

Otro rio llamado por sobrenombre de los Paeces en la Provincia y Gobierno de Buenos Ayres corre al O, y entra en el de Jacegua, entre el de Joseph Diaz y el paso del Chileno.

Otro de la Provincia y Gobierno del Chocó en el Nuevo Reyno de Granada nace de una laguna, y poco despues se une con el de Quito que nace de otra, y juntos forman el de Atrato.

Una Isla situada en el estrecho de Magallanes, cerca de la Costa del E, delante del Cabo de Monmouth.

Otra Isla pequeña de la mar del Sur en la Bahia de Panamá, delante del golfo de San Miguel.

PAC, Laguna pequeña de la Provincia y Gobierno de Yucatán.

un rio pequeño de la Provincia y Gobierno de Guayana ó Nueva Andalucía, nace en el país de los Indios Caribes feroces, y entra en el Caroni poco despues del parage donde le entra el caudaloso Arui.

PACABARA, Rio de la Provincia y Gobierno de Moxos en el Reyno de Quito, corre al N, y entra en el de Beni.

PACAIPAMPA, Pueblo de la Provincia y Corregimiento de Piura en el Perú, anexo al Curato de Frias.

PACÀJAS, Rio pequeño del país de las Amazonas, corre al N, entre los de Jacunda y Guanapú, y entra en el Marañon en el brazo que forma la Isla de Joanes; da el nombre este rio á una nacion de Indios que está poco conocida, y habita en la orilla boreal del Marañon, casi 80 leguas mas arriba del Paranaiba.

PACAJES, Provincia y Corregimiento del Perú, confina con la de Chucuito por el NO, por el N con la gran laguna Titicaca, por el NE con la Provincia de Omasuyos, y si-

PACAJES, Provincia y Corregimiento del Perú, confina con la de Chucuilo por el NO, por el N con la gran laguna Titicaca, por el NE con la Provincia de Omasuyos, y siguiendo por el E con la Ciudad de la Paz y Provincia de Cicasica, por el SE con el Corregimiento de Oruro y Provincia de Paria, por el S con la de Carangas, por el SO y O con la jurisdicción de la de Arica, mediando la cordillera..

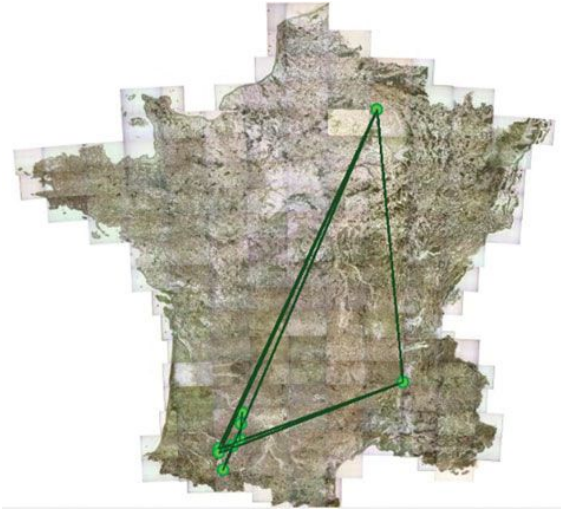
Pipeline adapté selon les niveaux de structuration des textes

Extrait sur la Champagne et le Sud-Ouest :

“*Sa puissance émotive, qui fut grande, demeura intacte jusqu'à la fin. Elle partit, le 28, de* ***Ruffec****, dans les circonstances qu'on connaît. Vers l'Ouest, elle gagna les forêts de* ***Chizé*** *et d'****Aulnay****, semble-t-il, à moins que celles-ci n'aient constitué un centre d'émotion locale.* ”

(Paris et al 2017)
https://dx.doi.org/10.1080/15420353.2017.1307306

La dimension spatiale de « La grande peur de 1789 » de Georges Lefebvre



Moins de structure, le cas de noms de lieux est faisable, néanmoins complexe au delà des EN “classiques : relations spatiales, co-références, …

*Subvention DAHN (Dispositif de soutien à l'archivistique et aux humanités numériques) par le MESRI.Travail réalisé avec Frédérique Mélanie-Becquet (LATTICE)

Une chaîne de traitement et d'analyse : annuaires PTM

Transcription
*

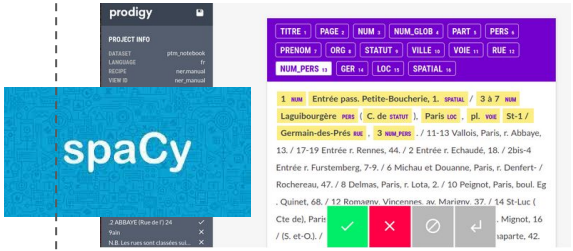
vérification

Structuration

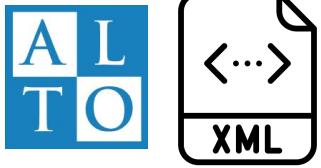
vérification

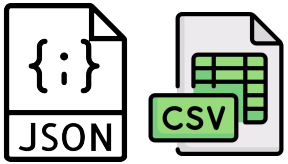
Analyses

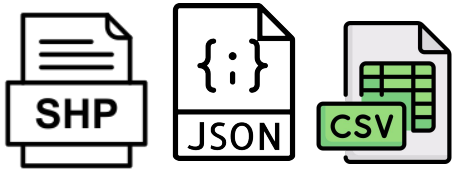
Outils



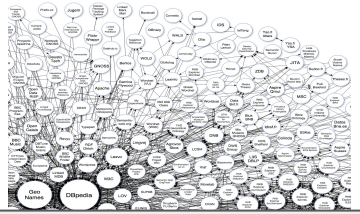
Formats







Chaîne TAL pour identifier les EN dans des textes littéraires (structuration)



http://fr.dbpedia.org/resource/Gare_du_Musée_d'Orsay

….
….
Gare d'Orsay
….
….
….

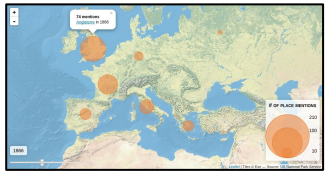
REN

Reconnaissance des entités nommées

NEL

Désambiguïsation et liage des entités nommées

….
…
Gare d'Orsay
….



Modes d'exploration de corpus textuels sous l'angle des EN

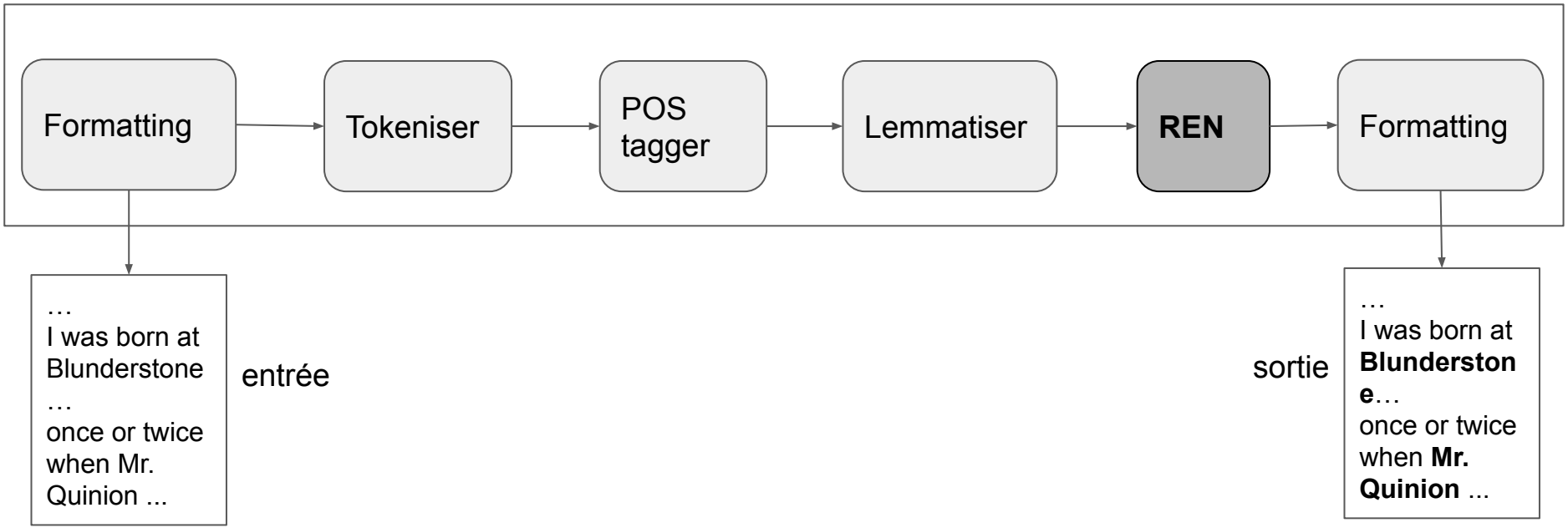
Enjeux EN et textes anciens : ambiguietés des localisations

- imprécision, (“Western Europe")
- subjectivité (“the main cities of Europe")
- noms vernaculaires (“Old World" pour l’America)
- référents changeant de frontières au fil du temps (Allemagne 1850 vs.1950)
- des lieux historiques qui ont cessé d'exister (“Gaule")
- ou qui ont changé de nom (“Place de Greve", aujourd’hui place de l’hôtel de ville)

… et plus largement, des référentiels “bruités”, peu exhaustifs, des représentations géométriques insuffisantes (ex: pour la Seine)

Exemples issus de : L'Heresiarque et cie (1910) de Guillaume Apollinaire, (23 nouvelles, 54166 mots, fiction) et “Qu'est-ce qu'une nation" (1882) d’Ernest Renan, (8500 mots, essai), corpus OBVIL

Un pipeline REN typique



Plusieurs configurations sont possibles et, en fonction de la langue, certains composants du pipeline peuvent être optionnels, à l'exception de la tokenisation. La segmentation des phrases peut être un autre composant.

Annuaires (PTM)

Q.4 QUATRE-VENTS (Rue des) [22]

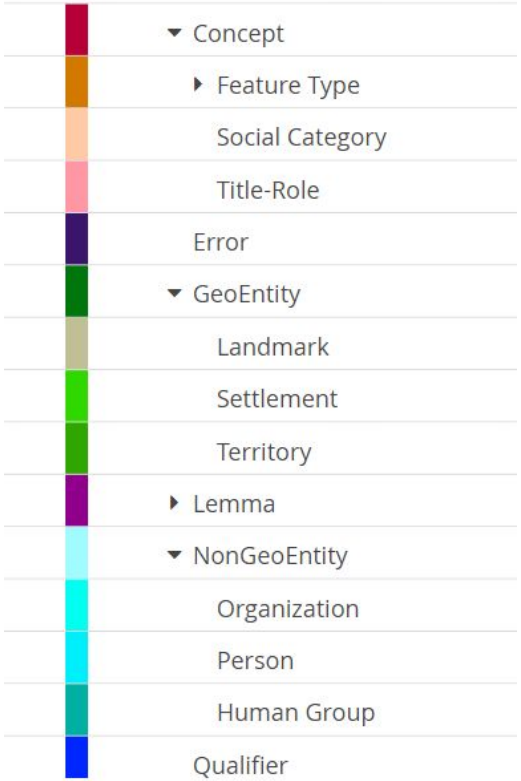
1 Entrée r. Condé, 2.
3 Burnel, Paris, r. Bourdonnais, 36.
5 Carreau, Paris, r. St-Denis, 270.
7 Tiers, Paris, r. St-Honoré, 91.
9 Entrée r. St-Sulpice, 6.
11 Entrée r. St-Sulpice, 8.
13 Salantin, Paris, r. Quatre-Vents, 13.
15 Prévost, Paris, r. Belzunce, 10.
17 Morat, Paris, pl. Denfert-Rochereau, 6.
19 Lefèvre, Paris, r. Vaugirard, 150.
2 Entrée Carrefour Odéon, 14.
4 Favre, Valcus-Charnet près Macon (S.-et-L.).
6 Guibert, Juvisy (S.-et-O.).
8 Cleray, Paris, r. St-Pétersbourg, 11bis.

étiquette	description
TITRE	Le titre de la microstructure
NUM	Numéro d'immeuble de la propriété
NUM_GLOB	Numéro (ou ensemble de numéros) référant à un même propriétaire
PART	Part de propriété (constr, terr, usufruit, nue-prop etc) qui est indiqué avant le nom du propriétaire
PERS	Nom du propriétaire
PRENOM	Prénom
ORG	Nom d'une organisation
STATUT	Vient après le nom, peut être Vve, héritiers, consorts,
VILLE	Ville du domicile.
VOIE	Nom de la voie (rue, rue de, boul etc) de l'adresse personnelle du propriétaire
RUE	Nom de la rue de l'adresse personnelle du propriétaire
NUM_PERS	Numéro d'immeuble de l'adresse perso du propriétaire
GER	pour Gérant et autres, peut contenir Nom et son adresse(rarement)
LOC	Département (si hors Paris) ou pays
SPATIAL	renvoie vers l'entrée sur une autre adresse

Q.4 QUATRE-VENTS (Rue des) 22

1 Entrée r. Condé, 2. / 3 Burnel, Paris, r. Bourdonnais, 36. / 5 Carreau, Paris, r. St-Denis, 270. / 7 Tiers, Paris, r. St-Honoré, 91. / 9 Entrée r. St-Sulpice, 6. / 11 Entrée r. St-Sulpice, 8. / 13 Salantin, Paris, r. Quatre-Vents, 13. / 15 Prévost, Paris, r. Belzunce, 10. / 17 Morat, Paris, pl. Denfert-Rochereau, 6. / 19 Lefèvre, Paris, r. Vaugirard, 150. /

Dictionnaire d'Alcedo : catégories à 2 niveaux



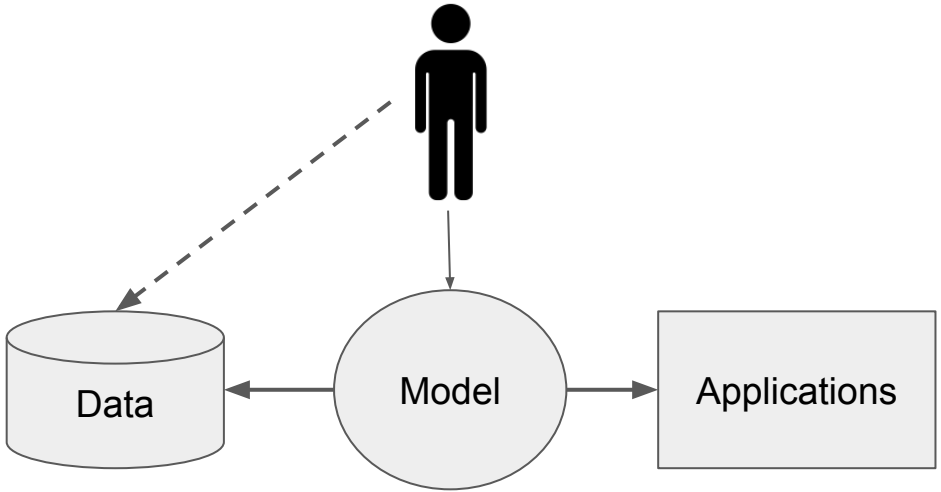
§§§id_00442§§§

ALLAUCA, Pueblo de la Provincia y Corregimiento de Yauyos, en el Perú, anexo al Curato de Tauripampa.

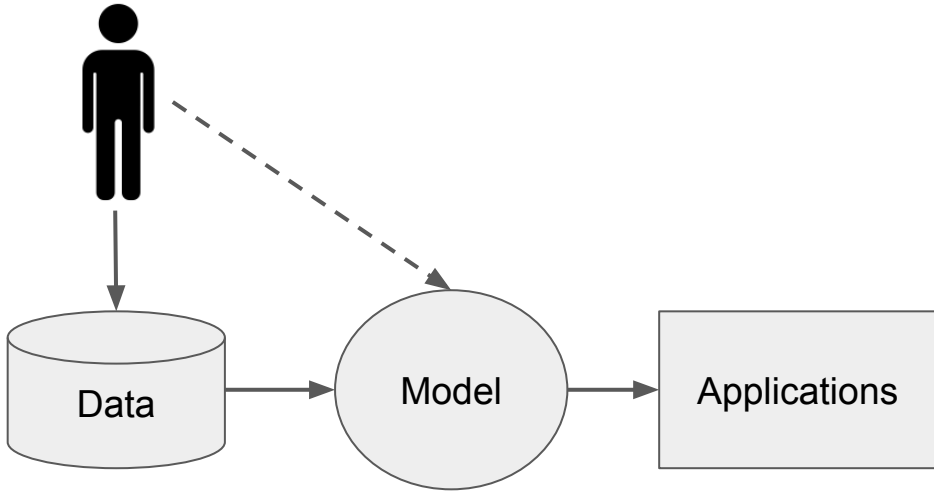
§§§id_00443§§§

ALLCA, Provincia antigua del Reyno del Perú, al Poniente del Cuzco: estos Indios bárbaros hicieron una grande y porfiada resistencia á Manco Capac, IV. Emperador de los Incas, llamado el Conquistador, favorecidos de la aspereza del terreno, que abunda de bosques, montes y lagos, como también de minas de oro y plata.

De systèmes à base de règles aux approches fondées sur les statistiques et les données



Symbolic systems



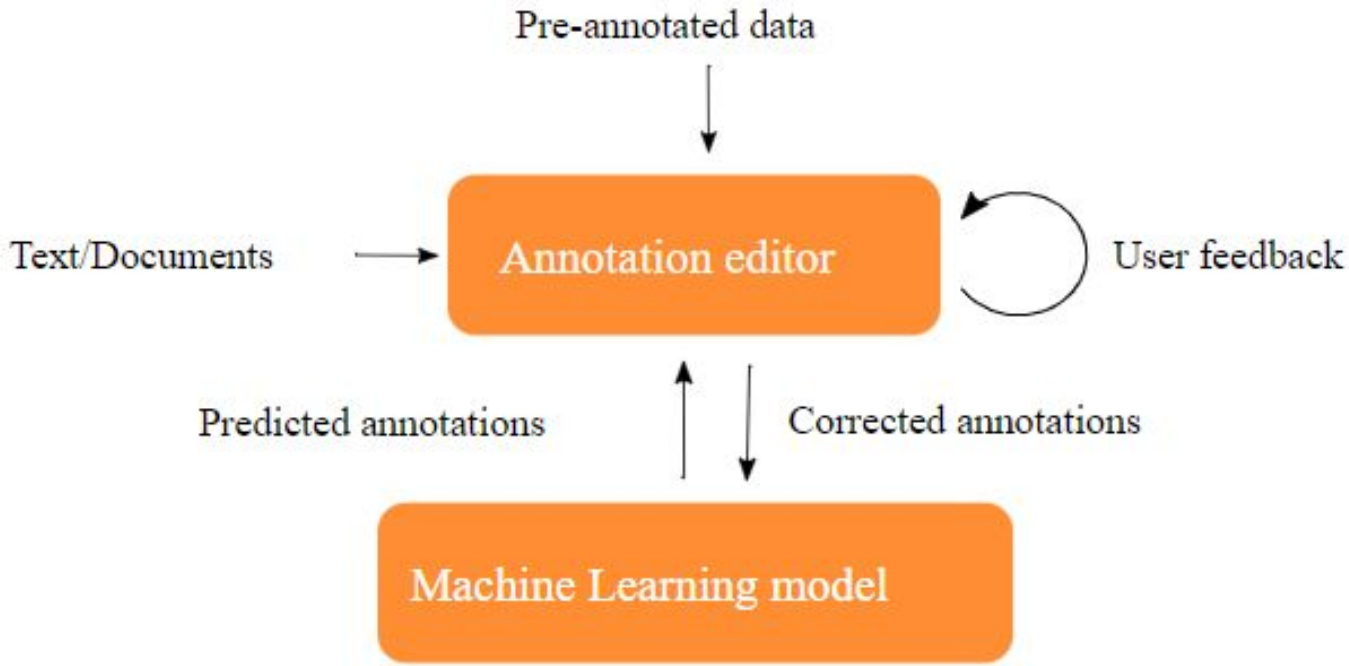
Statistical and data-driven approaches

interacts

visualise, evaluate, configure

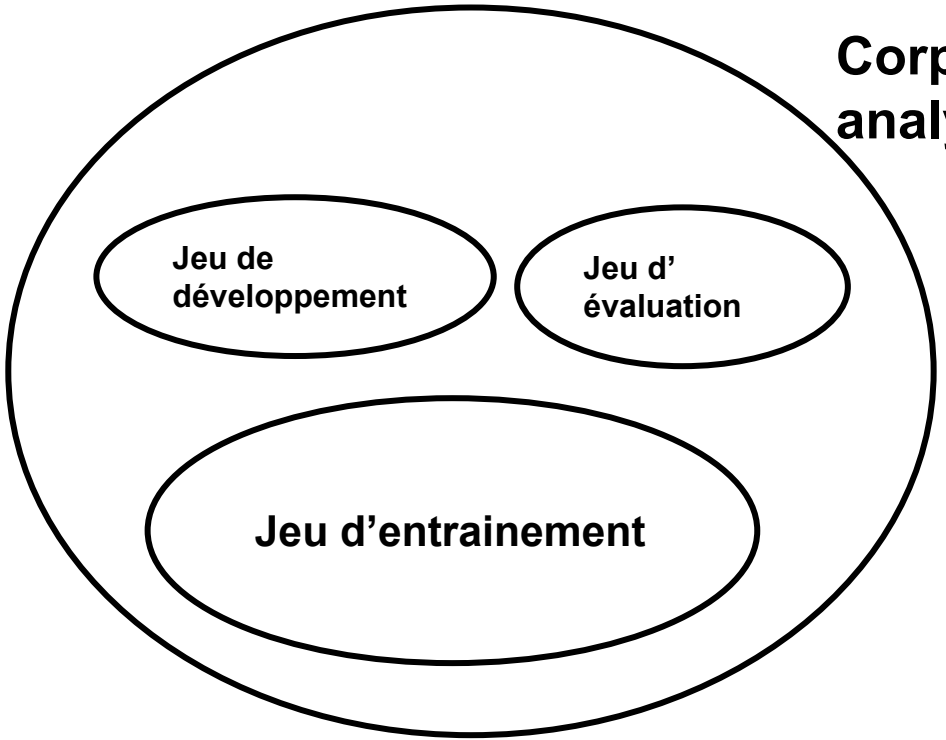
(Nouvel et al 2015)

Approche itérative

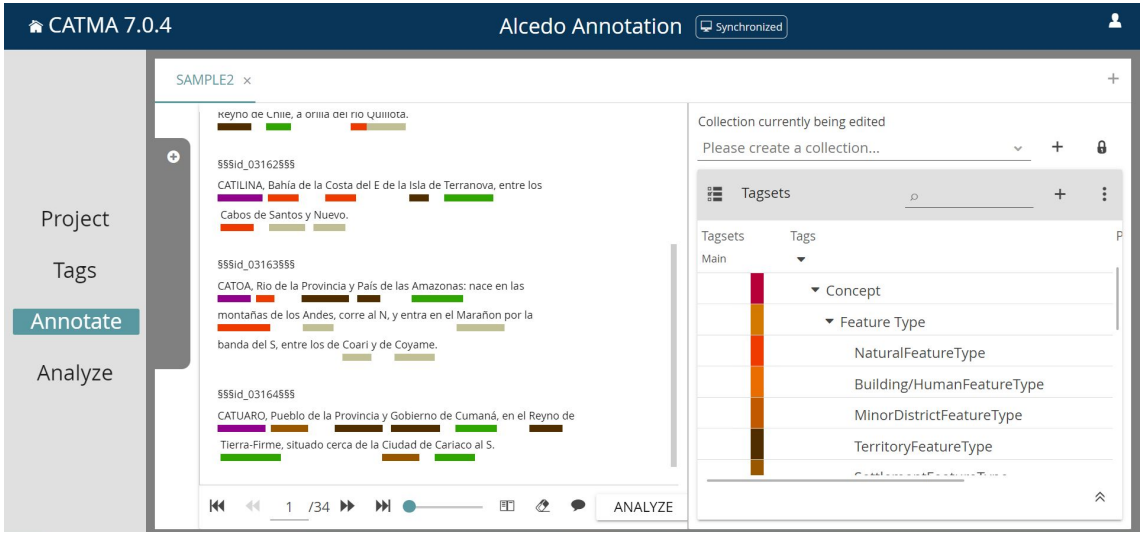


spaCy

Adaptation d'un système REN : apprentissage à partir d'un corpus annoté par un utilisateur

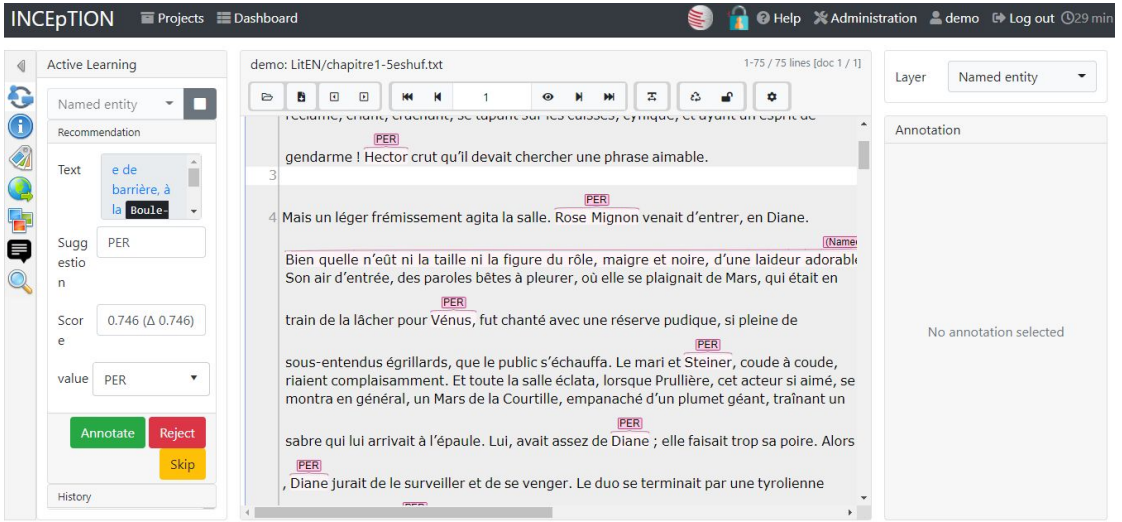


Création du jeu annoté manuellement pour l'entraînement d'un modèle



Catma:
https://app.catma.de/seven/

Inception:
https://inception-project.github.io/



Evaluation d'un système REN : annuaires PTM

Performance de l'annotation automatique à partir du modèle NER entraîné.

P (précision)
R (rappel)
F (f-score)

	P	R	F
NUM	98.68	100.00	99.34
PERS	98.61	99.65	99.12
VILLE	98.96	98.96	98.96
VOIE	99.63	100.00	99.81
RUE	100.00	100.00	100.00
NUM_PERS	100.00	99.62	99.81
LOC	100.00	100.00	100.00
PRENOM	100.00	60.00	75.00
GER	100.00	100.00	100.00
SPATIAL	100.00	98.75	99.37
STATUT	96.67	98.31	97.48
ORG	95.65	88.00	91.67
TITRE	100.00	100.00	100.00
PAGE	100.00	100.00	100.00
NUM_GLOB	100.00	33.33	50.00
PART	100.00	75.00	85.71

Travail réalisé avec Frédérique Mélanie-Becquet (LATTICE)

Avec Joanna Byszuk, Francesca Frontini, Ioana Galleron, Diana Santos, Ranka Stankovic
CLARIN Conference 2020,
https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf
https://github.com/COST-ELTeC

Alrahabi et al, 2021, DOI :
https://doi.org/10.4000/revuehn.1079
https://github.com/cvbrandoe/ArabicLitTAL

	Mesure	Partielle	Exacte
Farasa	Précision	**0,54**	**0,42**
	Rappel	0,45	0,35
	F1	0,49	0,38
Stanza	Précision	0,51	0,41
	Rappel	0,72	0,58
	F1	**0,59**	**0,48**
Madamira	Précision	0,35	0,25
	Rappel	**0,55**	**0,39**
	F1	0,43	0,31

https://github.com/OBVIL/Entites-nommees
(préliminaires, Zola, evaluation L3I à venir)

	F1 (exacte)	F1 (partielle)
Stanza	**0,69**	**0,55**
Spacy	0,65	0,40

	Cat	Precision	Recall
SEM-fra	LOC	0.465	0.395
	PERS	0.416	0.138
SPACY-fra	LOC	0.180	0.569
	PERS	0.526	0.629
PALAVRAS-por1	LOC	0.835	0.780
	PERS	0.905	0.901
SPACY-por1	LOC	0.338	0.728
	PERS	0.554	0.645
PALAVRAS-por2	LOC	0.624	0.693
	PERS	0.750	0.866
SPACY-por2	LOC	0.284	0.734
	PERS	0.591	0.707
Stanford-eng	LOC	0.438	0.495
	PERS	0.619	0.548
SPACY-eng	LOC	0.366	0.495
	PERS	0.691	0.453
SrpNER-srp	LOC	0.849	0.578
	PERS	0.820	0.729
SPACY-srp	LOC	0.354	0.308
	PERS	0.637	0.561

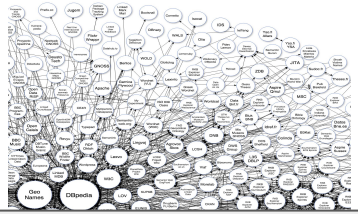
Évaluer des systèmes REN prêts à utiliser pour mesurer leur faiblesses avec l'idée de créer un benchmark.

Approches statistiques et fondées sur les données

La manière dont les données sont soumises au système est cruciale car :

- la **quantité** et la **qualité** des données peuvent rendre le système plus ou moins précis (peu de fausses correspondances), complet (peu de correspondances manquées), robuste (résistance au bruit),
- Le **type** de texte sur lequel l'entraînement est effectué conditionne l'applicabilité du modèle à d'autres types de texte,
- Le **prétraitement** (tokenisation, ...) et la façon dont les entités nommées sont définies influencent la façon dont elles sont reconnues.

Chaîne TAL pour identifier les EN dans des textes littéraires en vue d'une étude outillée linguistique et spatiale



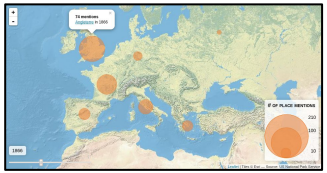
http://fr.dbpedia.org/resource/Gare_du_Musée_d'Orsay

….
….
Gare d'Orsay
….
….
….

REN

NEL

….
…
Gare d'Orsay
….



Reconnaissance des entités nommées

Désambiguïsation et liage des entités nommées

Modes d'exploration de corpus textuels sous l'angle des EN

Lier les EN aux référentiels

Named Entity Linking, il s'agit d'un moyen d'établir un lien explicite entre les mentions d'EN cités dans le texte et les objets du monde auxquels elles réfèrent. Créer ce lien est important afin de lever toute ambiguïté sur l'identité de l'EN. Il y a donc deux objectifs :

- ❏ Désambiguiser

"Goncourt" - Edmond de Goncourt ou Jules de Goncourt ?

"Voltaire", "François-Marie Arouet" - manières de désigner la même personne

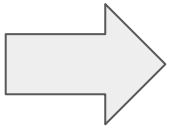
- ❏ Lier à un référentiel

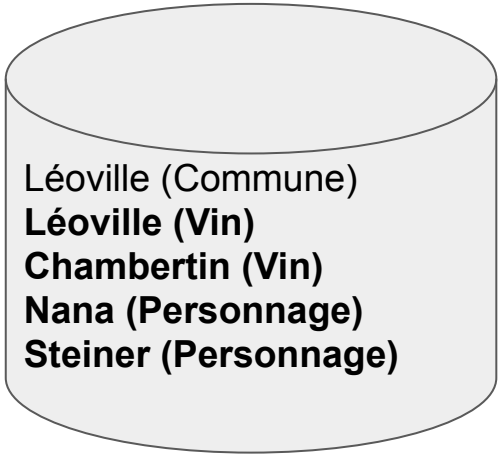
"Voltaire" - associer l'entrée correspondante dans Wikidata (https://www.wikidata.org/wiki/Q9068)

Lier les EN aux référentiels

Léoville ou **Chambertin** ? murmura un garçon, en allongeant la tête entre **Nana** et **Steiner**, au moment où celui-ci parlait bas à la jeune femme.

Texte annoté en EN :
Nana (Zola)





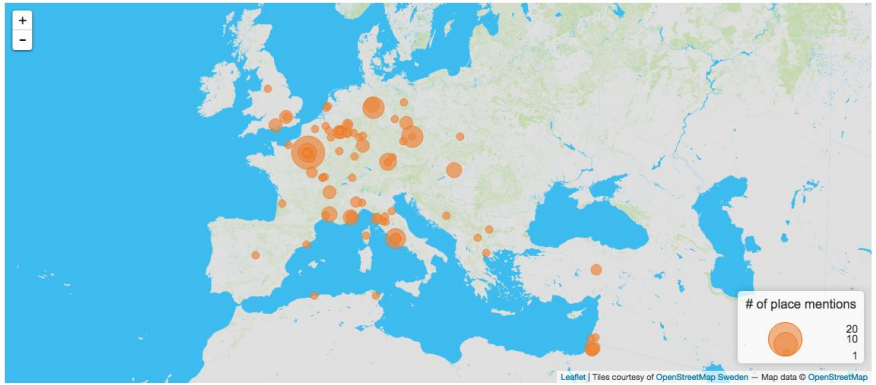
Référents possibles

La mention étudiée hors de son contexte est ambiguë. Un humain peut typiquement lever l'ambiguïté par utilisation conjointe de **la connaissance du monde** dont il dispose et **d'indices contextuels**.

L’empreinte spatiale d’un texte pour rendre compte de la géographie d’un auteur et les échelles mobilisées

Visualization

The following map shows distribution of places mentioned in the input TEI-XML file, geo-coordinates are obtained via French DBpedia.



140 places are displayed on the map.

19 places were not included on the map because geo-coordinates were unavailable, these are: Rhin, Hambourg, Queensland, Bohême, La Nouvelle-Orléans, Berlin, Amsterdam, Neckar, Ile-de-France, Bavière, montagnes Rocheuses, Provence, Monte-Carlo, empire des Habsbourg, Moldau, Danube, Europe, royaume de Juda, Savoie

Guillaume Apollinaire’s « Le passant de Prague »

Visualization

The following map shows distribution of places mentioned in the input TEI-XML file, geo-coordinates are obtained via French DBpedia.

137 places are displayed on the map.

17 places were not included on the map because geo-coordinates were unavailable, these are: Berlin, Rhin, Bohême, montagnes Rocheuses, Queensland, royaume de Juda, La Nouvelle-Orléans, Provence, Neckar, empire des Habsbourg, Danube, Moldau, Ile-de-France, Hambourg, Bavière, Savoie, Amsterdam

You can download the resulting annotated XML-TEI file here

Quelques remarques finales : enjeux EN et corpus des HN

- **langues de textes** diverses, anciennes, melangées, non normalisés
- **catégories EN** fines vs. trop générales ; différentes **définitions des EN** selon les cas d'applications (ex: édition TEI)
- Intégrer dans les modèles REN :
  - des éléments de la **mise en forme** du document imprimé (travaux de Joseph Chazalon et l'EPITA et l'IGN sur les annuaires dits Bottin)
  - **bruit** présente dans le texte issu de la numérisation (voir Séminaires du Groupe NER for historical documents https://ner-for-historical-docs.github.io/)
- **référentiels** de données anciennes peu disponibles pour lier les EN
- **évaluation** des performances des systèmes : manque de jeux de tests pour comparer
- **Modèles à adapter**, fine-tuning, modèles génératifs, besoin de ressources de calcul
- **Au delà des EN**: annoter les vocabulaires contrôlés et expressions polylexicales reste simple, les relations spatiales faisable dans la plupart des cas, mais peu de propositions pour les entités non nommées et les co-références présentes dans des textes littéraires (sauf pour les travaux en cours du laboratoire LATTICE sur le French BookNLP)

Ressources pour découvrir le REN

Spacy - modèles disponibles : https://spacy.io/models

Quelques notebooks Python pour démarrer :

1) https://github.com/cvbrandoe/coursTAL/tree/master/2023/notebooks :
   coursTAL_EltecFRL2_StepA_1eAnnotStanza_CorrTagTog.ipynb
   coursTAL_EltecFRL2_StepB_TrainSpacy.ipynb
   coursTAL_EltecFRL2_StepC_UseTrModelSpacy.ipynb
2) https://github.com/ludovicmoncla/tutoriel-geoparsing
3) https://github.com/orgs/Consortium-ARIANE/
   (issus de la formation ARIANE à Lyon, ils seront déposés ici la mi-novembre 2023)

MERCI

Evaluation d'un système REN

gold: ***Phébus*** *parut en* ***Postillon de Lonjumeau*** *et* ***Minerve*** *en Nourrice normande.*

test: ***Phébus*** *parut en Postillon de* ***Lonjumeau*** *et* ***Minerve*** *en* ***Nourrice*** *normande.*

True positive (TP) False Negative (FN) TP False positive (FP)

- **Recall** : the number of correctly annotated entities wrt the total of manually annotated entities in the gold
  = TP / (TP + FN)
- **Precision** : the number of correctly annotated entities wrt the total of returned entities
  = TP / (TP + FP)
- **F-score**, harmonic mean between recall and precision.

It is also possible to **extend** these measures to use **relaxed match** when comparing annotations, instead of **strict match** as presented above.